DOCUMENT RESUME

ED 307 299 TM 013 318

AUTHOR Seaman, Michael A.; And Others
TITLE New, Improved Multiple-Comparison Procedures: More Pep with Each Step.
PUB DATE Mar 89
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Algorithms; *Comparative Analysis; *Error of Measurement; *Monte Carlo Methods
IDENTIFIERS Bonferroni Procedure; *Multiple Comparisons; Paired Comparisons; Stepwise Regression; Tukey Statistic; *Type I Errors

ABSTRACT

This Monte Carlo investigation provides some possible solutions to problems related to choosing multiple-comparison methods that maximize true rejections and minimize false ones. It has been argued that the traditional Bonferroni approach to multiple comparisons, which satisfies the statistician's family-wise Type I error concerns, could be improved to satisfy the researcher's power concerns by converting it from a simultaneous procedure to a sequential (or stepwise) procedure. In this study, the algorithm proposed by W. H. Press et. al. (1986) was implemented to obtain random normal deviates. This technique generates random uniformly distributed values and applies the Box-Muller transformation to them. The problem of all-possible pairwise comparisons in a set of "K" independent means was addressed to: (1) assess the Type I error and power characteristics of the sequential Bonferroni approaches on the basis of an empirical study; and (2) compare those approaches with both the traditional Bonferroni and Tukey multiple-comparison procedures. Seven multiple-comparison procedures were used for testing the pairwise comparisons among means. Focus was on the increase in power that can be obtained by sequential tests when the Type I error is to be controlled for a family of comparisons. To provide a broad comparison of both Type I error rates and power among the various procedures, several parameters were varied. Each test was performed using the algorithm prescribed by the particular multiple-comparison procedure based on a family-wise alpha of 0.05. Results suggest that there are valid reasons for replacing the traditional (simultaneous) Bonferroni and Tukey procedures with improved sequential tests. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307299

New, Improved Multiple-Comparison Procedures: More Pep With Each Step

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Michael A. Seaman
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Michael A. Seaman

Todd M. Franke

Ronald C. Serlin

Joel R. Levin

University of Wisconsin-Madison

March, 1989

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco

TM013318

BEST COPY AVAILABLE

NEW, IMPROVED MULTIPLE-COMPARISON PROCEDURES: MORE PEP WITH EACH STEP

The problem of choosing appropriate statistical methods for the multiple comparisons of means has a long and somewhat perplexing history. Of the numerous options that have been proposed over the last forty years, the question still remains: Which procedure is best? On the one hand, statisticians are typically attracted to procedures that maintain a reasonable and dependable Type I error probability--which, for the multiple-comparison problem, has come to be accepted as the familywise α. On the other hand, educational researchers in search of mean differences occasionally select procedures on the basis of their smaller critical values (and, presumably, greater statistical power), ignoring issues of the potentially inflated familywise α associated with such procedures. What is needed, naturally, are multiple-comparison procedures that both maximize true rejections and minimize false ones. The present Monte Carlo investigation offers some possible solutions to this quandary.

As the name suggests, the term familywise error rate implies that there is more than one comparison to be made, not all necessarily independent and comprising subsets of a larger single family of effects. The most common example of this occurs in the post-hoc or planned analysis of differences among means which could, or may, have been analyzed as an omnibus test of the null hypothesis that all contrasts among means are zero. A familywise Type I error occurs when at least one null hypothesis in the family of hypotheses is falsely rejected. Setting the familywise Type I error rate at α ensures that the

probability of committing no Type I errors in the complete family of hypothesis tests is at least 1-α.

Commonly, the familywise error rate is held below α through the use of a simultaneous procedure based on either the Bonferroni inequality or the Tukey range statistic. The Bonferroni test (Dunn, 1961) holds the maximum Type I error rate at α by testing each contrast with a Type I error rate of α/C, where C is the number of contrasts. The simultaneous Tukey test (Kirk, 1982), denoted the Tukey HSD, for honestly significant difference, uses percentiles of the distribution of the range statistic as critical values. The range statistic maintains the familywise Type I error rate if the contrasts being tested are pairwise.

The immediate impetus for the study was provided by the work of Holm (1979) and Shaffer (1986). What Holm did was to argue mathematically that the traditional Bonferroni approach to multiple comparisons, which satisfies the statistician's familywise α concerns, could be improved to satisfy the researcher's power concerns by converting it from a simultaneous procedure to a sequential (or stepwise) procedure. Briefly, Holm proved that for a set of C planned comparisons the familywise Type I error probability will not exceed α if: 1) one tests the largest mean difference based on a Type I error probability of α/C; 2) the next largest mean difference is tested based on $\alpha/(C\text{-}1)$, and so on, up to the C-th comparison, whose test is based on α; and 3) on any step the procedure is terminated if a nonsignificant difference is encountered. The basic logic rests on the fact that at any step when a rejection occurs either a Type I error was committed or a correct rejection was made. If a Type I error occurs, it can be

considered a familywise error and thus any errors made in subsequent steps are irrelevant. On the other hand, if the rejection was valid, then α can be divided among the remaining contrasts.

Shaffer extended this argument by modifying Holm's sequential approach (viz., by incorporating implied mean inequalities that logically follow from rejections that occur during the sequential-testing procedure) to make it even more powerful. In particular, at each step α is divided by the number of possible true null hypotheses remaining (a number that is a function of K, the number of means included, and the pattern of rejections). An example is provided in Table 1 ($K = 4$, $n = 7$ observations per group, familywise $\alpha = .05$) to compare the stepwise critical values associated with Holm's and Shaffer's approaches.

Since the maximum number of true hypotheses possible at any step depends upon which contrasts were previously rejected, the sequence of critical values is not unique. One possible sequence for the $K = 4$ case is provided in the example. Here the ordered set of means (say (1, 2, 3, 4)) was tested in a stepwise fashion. At the start, there are six pairwise comparisons, which is also the maximum number of true hypotheses. Rejection of the test of the largest contrast, 1 versus 4, reduces the number of possibly true hypotheses to those within the set (1, 2, 3) or the set (2, 3, 4). There are a maximum of three true hypotheses, since the hypotheses within only one of these sets, but not both, could be true. If 1 versus 3 is now rejected, the sets to be tested are (1, 2), and (2, 3, 4). Notice that the 2 versus 3 comparison is contained in the second set. The pattern in the example assumes that 1 versus 2 was now the largest contrast, so if this is rejected, the

remaining set is (2, 3, 4), and there are still three true hypotheses possible. The remaining two steps proceed in this fashion, with only one possible true hypothesis existing for each of the steps after a rejection of 2 versus 4.

Sequential Bonferroni procedures have a great deal to offer the educational researcher, especially in light of the fact that currently popular sequential procedures, such as those proposed by Fisher (1935), Duncan (1955), and Newman (1939) and Keuls (1952), do not preserve their nominal familywise αs with all patterns of mean differences for particular numbers of means (Carmer & Swanson, 1973).

The test proposed by Fisher, denoted the LSD test (for least significant difference), has as its first step an omnibus F test of equality among all means. Following a significant finding, multiple t tests are conducted on all pairwise differences.

The Newman-Keuls range test is a sequential procedure which begins by comparing the largest difference between means to the same critical value used by the Tukey HSD. Thus, for the ordered means (1, 2, 3, 4) a test of 1 versus 4 is conducted as it would be using the HSD method. Following a rejection of this difference, means which are 3 steps apart (i.e., 1 versus 3 and 2 versus 4) are tested using the Tukey critical values for the pairwise comparisons among 3 means. Similarly, comparisons between means which are 2 steps apart are compared to Tukey critical values for 2 means. Duncan's test follows the same procedure as the Newman-Keuls test, except that critical values are not range statistics but, rather, are values calculated by Duncan. Use of these critical values ensures that the maximum familywise Type I error rate is equal to $(K\text{-}1)\alpha$.

While it is recognized that sequential procedures exist which do meet the criteria of controlling the familywise Type I error rate (see Ryan, 1960; Einot & Gabriel, 1975; Welsch, 1977), these procedures depend on testing the range and *F* statistics against critical values which are not commonly tabled. Thus, the potential use of such procedures for the educational researcher is limited and so these tests are not examined here.

The present study addressed the problem of all-possible pairwise comparisons in a set of *K* independent means and had two major purposes: 1) to assess the Type I error and power characteristics of the sequential Bonferroni approaches on the basis of an empirical study, and 2) to compare those approaches with both the traditional Bonferroni and Tukey multiple-comparison procedures. The rationale for the latter plan was as follows: Tukey's simultaneous procedure is both sound in its familywise α control and specifically suited to the all-pairwise multiple-comparison problem. Indeed, it is well established that when all pairwise comparisons are of interest, the simultaneous Tukey procedure is superior to the simultaneous Bonferroni approach in terms of statistical power. But how do the sequential Bonferroni tests compare with the simultaneous Tukey procedure? How would the various methods compare under different configurations of means, mean differences, and sample sizes? Can any general recommendations be made?

Method

In our Monte Carlo study, the algorithm proposed by Press, Flannery, Teukolsky, and Vetterling (1986) was implemented to obtain random normal deviates. This technique generates random uniformly

distributed values and then applies the Box-Muller (1958) transformation to them.

Seven multiple-comparison procedures were used for testing the pairwise comparisons among means: simultaneous Bonferroni (B), Tukey's HSD (HSD), Scheffe's S-Method (S), Fisher's LSD (LSD), Newman-Keuls sequential range (NK), Holm's sequential Bonferroni (SB), and Shaffer's modified sequential Bonferroni (MSB). Of major interest is the increase in power which can be obtained by sequential tests when the Type I error is to be controlled for a family of comparisons.

To provide a broad comparison of both Type I error rates and power among the various procedures, several parameters were varied: K (3,4,5, nd 6), n (small: 5-9, medium: 13-21, and large: 25-41), and the pattern of partial non-nullity (two, three, four, or five means equal). The common population variance was chosen as 1.0 for each simulated experiment. With the exception of the complete-null case (i.e., all pairs of means equal), differences between means were chosen such that the power associated with the largest difference was .80 for the simultaneous Bonferroni procedure. Values of the population standardized mean difference corresponding to this power were obtained using the noncentral F distribution. For the $K > 3$ layouts with more than one inequality in means, the remaining differences were chosen to be half of the largest standardized difference.

Each test was performed using the algorithm prescribed by the particular multiple-comparison procedure based on a familywise α of .05. For each combination of mean configuration, K, and n, investigated, 5000 samples were drawn. Empirical familywise Type I error rates were tallied, along with two types of power. One type (Power A) was the

empirical estimation of power for each pairwise difference in which the alternative hypothesis is true. The second type (Power B) was the percentage of true rejections associated with each procedure for each step of the testing process. Thus, the percentage of correct rejections on the largest difference was calculated (no matter which contrast this happened to be), followed by the percentage of correct rejections on the next largest difference, and so on.

Since the actual Type I error rate in this study for each of the procedures compared is no greater than .05, the width of the 95 percent acceptance region for a test of the hypothesis that the true proportion is .05 is approximately .012. Thus, the estimated value will usually be within .006 of the actual α.

Results and Discussion

Type I Error Comparisons

For the complete-null cases, all seven procedures maintained the familywise αs. They also did in the $K = 3$ partial-null case. However, in the $K = 4$, 5, and 6 partial-null cases it is known that the LSD and NK will not always hold the Type I error rate less than or equal to α (Arnold, 1981). For these cases, examining the power of the LSD and NK procedures is inappropriate, and comparisons between these two tests and the remaining procedures are meaningless. The LSD and NK tests were included in the study in order that power comparisons in the common $K = 3$ situation (when the familywise Type I error rate is maintained at α) could be made.

Type I error rates averaged across sample size are reported in Table 2 for each of the seven tests for $K = 3$ and for the five valid

tests for $K > 3$. In no instance is any Type I error rate outside of expected bounds. Estimates of α for the complete null cases and the cases with K-1 means equal to each other and different from the other mean are shown. The Type I error rate results for the remaining partial-null cases are similar and therefore not shown.

For the complete null cases, the simultaneous and sequential procedures based on the Bonferroni inequality (B, SB, MSB) will necessarily have the same Type I error rate. This is because the only comparison of interest is the contrast between the smallest and largest means, and the critical value used for this test is the same in all three procedures. Similarly, both procedures based on the studentized range (HSD, NK) have their own identical Type I error rate in the complete null cases.

Power Comparisons

Table 3 shows some of the Power A comparisons for equivalent contrasts averaged across sample sizes. For example, in the case of $\mu_1 > \mu_2 = \mu_3$, the contrast μ_1-μ_2 has the same power as μ_1-μ_3, and so the estimates of power for these two contrasts were averaged. It was also appropriate to compute average power across the small, medium, and large sample sizes, since the differences between means for each of these samples were chosen so as to maintain constant power.

It is obvious that the power for a sequential test will show improved performance over the corresponding simultaneous test across all patterns of mean differences. The empirical results estimate the size of this difference. The SB showed power gains of two to three percent over the B. The MSB, in turn, resulted in two to four percent more power than the SB. The MSB, being the most powerful of the Bonferroni

methods, achieved as much as a six percent increase over the more commonly used B test procedure.

For the $K = 3$ case, when the Fisher LSD is appropriate, the LSD procedure is the most powerful test. There was approximately a two percent gain in power over the MSB, giving about an eight percent gain over the B. Though it should be emphasized that use of the LSD is limited to the three mean setting, the power of the test and the fact that the procedure only requires calculating the omnibus F statistic and subsequent t tests are strong arguments for its use when appropriate.

The Power B calculations revealed the size of the increase in power achieved by the sequential procedures over the simultaneous procedures after the first step. Table 4 shows Power B results for two three mean cases, $\mu_1 > \mu_2 = \mu_3$ and $\mu_1 > \mu_2 > \mu_3$. The latter of these is especially interesting, since it emphasizes that the gain of the sequential procedures over the simultaneous procedures can be considerable as the number of pairs with mean differences other than zero increases (i.e., more pairwise comparisons which should reject). This is the result of the ever-diminishing size of the critical values used.

General Conclusions and Recommendations

The results of the present study should be of interest to both educational statisticians and researchers. What we have shown is that there are valid reasons for replacing the traditional (simultaneous) Bonferroni and Tukey procedures with improved sequential tests. The new tests are easy to perform, being based on the same test statistic as the simultaneous tests and merely requiring changing the critical value at each step. In contrast to other proposed modifications of the additive

Bonferroni inequality that result in trivial power increases (e.g., Games, 1977), the power increases here (especially those resulting from Shaffer's modified sequential approach, whose critical values can be determined on a case-by-case basis) are substantial. Also, in contrast to many other currently popular multiple-comparison procedures, these new procedures increase power with no compensating familywise Type I error costs. Thus, whenever familywise (rather than per-contrast or per-hypothesis) αs are the focus, this new class of sequential procedures can be adopted with double-barreled confidence.

The choice of which procedure to use depends on the number of means being compared as well as the pattern of true or false null hypotheses anticipated by the researcher. For all pairwise comparisons of three means the choice is obvious: Fisher's LSD procedure is the most powerful. For pairwise comparison of more than three means, Shaffer's modified sequential test is most powerful, though this test requires evaluating sets of possible equalities between means at each step.

A further consideration is that the difference in power between the MSB and the simultaneous range test, HSD, can be small, depending on the number of means being compared and the size of the mean differences. When most of the differences between means are zero and the number of pairwise comparisons is relatively large, the difference between the MSB and HSD is slight. In practice, however, theory often predicts most or all of the comparisons to be nonzero differences. In these cases the increase in power which can be achieved by using the MSB may be considerable.

For the researcher who prefers a simple algorithm, as well as for instructors of statistics who wish to present students with a straightforward test procedure, the Tukey HSD and the sequential Holm procedure are the most powerful of the tests which can be classified in this category. For at least those patterns of means examined in this study, however, the HSD never exceeded the power of the MSB.

In any case, the sequential procedures provide a strong argument against automatic use of simultaneous test procedures when familywise Type I error protection is chosen. Extensions of the Bonferroni sequential procedures are many and varied. Applications could be found for use of the procedures when testing both complex, nonpairwise comparisons and certain multivariate tests. Tables of commonly used percentiles of the studentized range distribution and the *F* distribution are all that are needed to find critical values for all of the procedures studied here, so the usefulness of these tests should transcend theory and find its way into the hands of educational researchers.

References

Arnold, S. F. (1981). The theory of linear models and multivariate analysis. New York: John Wiley & Sons.

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. Annals of Mathematical Statistics, 28, 610-611.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 68, 66-74.

Duncan, D. B. (1955). Multiple-range and multiple-*F* tests. Biometrics, 11, 1-42.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. Journal of the American Statistical Association, 70, 574-583.

Fisher, R. A. (1935). The design of experiments. Edinburgh: Oliver and Boyd.

Games, P. A. (1977). An improved *t* table for simultaneous control on *g* contrasts. Journal of the American Statistical Association, 72, 531-534.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Keuls, M. (1952). The use of the studentized range in connection with an analysis of variance. Euphytica, 1, 112-122.

Kirk, R. E. (1982). Experimental design (2nd ed.). Belmont, CA: Brooks/Cole.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of independent estimate of a standard deviation. Biometrika, 31, 20-30.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). Numerical recipes: The art of scientific computing. Cambridge, MA: Cambridge University Press.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. Psychological Bulletin, 57, 318-328.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. Journal of the American Statistical Association, 72, 566-575.

Table 1

Example comparing Holm and Shaffer critical Bonferroni values for $K = 4$, $n = 7$, familywise $\alpha = .05$, and assuming that the hypotheses are tested in order from the largest to the smallest difference

Hypothesis	Holm Critical Value		Max. True Hyps.	Shaffer Critical Value	
1	$\alpha/6$	$t = 2.87$	6	$\alpha/6$	$t = 2.87$
2	$\alpha/5$	$t = 2.80$	3	$\alpha/3$	$t = 2.57$
3	$\alpha/4$	$t = 2.70$	3	$\alpha/3$	$t = 2.57$
4	$\alpha/3$	$t = 2.57$	3	$\alpha/3$	$t = 2.57$
5	$\alpha/2$	$t = 2.39$	1	α	$t = 2.06$
6	α	$t = 2.06$	1	α	$t = 2.06$

Table 2

Type I Error Rates (averaged across sample sizes)

All μs Equal

K =	3	4	5	6
B	.0437	.0386	.0387	.0397
SB	.0437	.0386	.0387	.0397
MSB	.0437	.0386	.0387	.0397
HSD	.0499	.0488	.0499	.0527
NK	.0499	-----	-----	-----
S	.0393	.0261	.0196	.0153
LSD	.0510	-----	-----	-----

$\mu_1 > K$-1 μs Equal

K =	3	4	5	6
B	.0151	.0212	.0253	.0270
SB	.0373	.0371	.0373	.0355
MSB	.0506	.0431	.0407	.0373
HSD	.0173	.0274	.0343	.0356
NK	.0506	-----	-----	-----
S	.0136	.0150	.0142	.0103
LSD	.0507	-----	-----	-----

Table 3

Type A Power Calculations (averaged across sample sizes and equivalent contrasts)

$\mu_1 > K$-1 Equal μs

K =	3	4	5	6
B	.7965	.7993	.8016	.7995
SB	.8262	.8268	.8240	.8185
MSB	.8624	.8540	.8447	.8350
HSD	.8141	.8234	.8309	.8319
NK	.8718	-----	-----	-----
S	.7841	.7586	.7253	.6819
LSD	.8801	-----	-----	-----

$\mu_1 > \mu_2 > K$-2 Equal Means (largest difference)

K =	3	4	5	6
B	.7912	.8026	.8000	.7996
SB	.7976	.8210	.8216	.8175
MSB	.8026	.8464	.8376	.8357
HSD	.8102	.8267	.8318	.8316
NK	.8204	-----	-----	-----
S	.7830	.7604	.7230	.6834
LSD	.8214	-----	-----	-----

$\mu_1 > \mu_2 > K$-2 Equal Means (smallest non-zero difference)

K =	3	4	5	6
B	.2222	.1842	.1638	.1475
SB	.2929	.2380	.2029	.1778
MSB	.3651	.2678	.2213	.1918
HSD	.2440	.2094	.1911	.1766
NK	.3682	-----	-----	-----
S	.2130	.1501	.1123	.0832
LSD	.3737	-----	-----	-----

Table 4

Type B Power Calculations for Three Means

$\mu_1 > \mu_2 = \mu_3$

	Step 1	Step 2	Step 3
B	.9073	.6843	.0015
SB	.9073	.7391	.0061
MSB	.9073	.8099	.0075
HSD	.9200	.7063	.0019
NK	.9200	.8161	.0075
S	.8989	.6679	.0014
LSD	.9263	.8265	.0075

$\mu_1 > \mu_2 > \mu_3$

	Step 1	Step 2	Step 3
B	.8046	.4058	.0252
SB	.8046	.4942	.0846
MSB	.8046	.6328	.0954
HSD	.8232	.4426	.0324
NK	.8232	.6384	.0952
S	.7964	.3902	.0224
LSD	.8238	.6496	.0954